High-Performance Light Field Reconstruction with Channel-wise and SAI-wise Attention

Zexi Hu[1], Yuk Ying Chung[1], Seid Miad Zandavi[1], Wanli Ouyang[1], Xiangjian He[2], and Yuefang Gao[3]

School of Computer Science, University of Sydney, Sydney, Australia
huzexi@outlook.com
{vera.chung, miad.zandavi, wanli.ouyang}@sydney.edu.au
School of Computing and Communications, University of Technology, Sydney, Australia
Xiangjian.He@uts.edu.au
College of Mathematics and Informatics, South China Agricultural University, Guangzhou, China
gaoyuefang@scau.edu.cn

Abstract. Light field (LF) images provide rich information and are suitable for high-level computer vision applications. To acquire capabilities of modeling the correlated information of LF, most of the previous methods have to stack several convolutional layers to improve the feature representation and result in heavy computation and large model sizes. In this paper, we propose channel-wise and SAI-wise attention modules to enhance the feature representation at a low cost. The channel-wise attention module helps to focus on important channels while the SAI-wise attention module guides the network to pay more attention to informative SAIs. The experimental results demonstrate that the baseline network can achieve better performance with the aid of the attention modules.

Keywords: light field · image processing · deep learning.

1 Introduction

Light field (LF) images can provide both angular and spatial information by capturing the appearance of objects from several angles in one shot. Compared with regular images, such a feature realizes many high-level computer vision applications such as depth extraction [5–7,9,14,15], refocusing [4,12] and material classification [16]. With the emergence of commercial and industrial LF cameras, *e.g.* Lytro and RayTrix, LF has drawn more attention. However, LF cameras suffer from the inherent trade-off between angular and spatial resolution caused by the limitation of sensor space.

LF reconstruction is adopted for alleviating this dilemma which focuses on boosting the angular resolution, *i.e.* the number of SAIs. For example, a densely sampled 8×8 LF is possible to be reconstructed from a sparsely sampled 2×2

LF. With the introduction of deep learning, the performance is significantly improved by the powerful feature representation learned from training samples. In [10], Kalantari *et al.* have proposed the first deep learning-based method to tackle this task by extracting disparity maps using a disparity network, and then warping the input SAIs into the novel SAIs by a color network. With the deep features, it has achieved state-of-the-art performance, nevertheless, it has suffered from intensive computation as it has to reconstruct SAIs separately. Yeung *et al.* [17] proposed a fully convolutional network (FCN) where the 4D LF image can be processed jointly. To mitigate the intensive computational problem of directly operating convolution on the 4D LF data, they proposed a Spatial-Angular Alternating convolutional layer to approximate the 4D convolution by a spatial convolution and an angular convolution. However, to acquire higher-level features and a larger receptive field, these methods have to stack more convolutional layers, which leads to heavy computation and large model size. On the other hand, they treat the features in different locations equally. It is possible that some of the learned features are more important and should not be paid the same attention as other features.

In this paper, we propose two different kinds of attention modules, namely channel-wise and SAI-wise attention for LF. We have assumed that in the processing of LF reconstruction, some of channels and SAIs may carry more important information and to enhance the feature representation these components should be reinforced while the other trivial components should be suppressed. To verify this hypothesis, we apply the two attention modules on a baseline network and propose a novel Channel-wise and SAI-wise Attention (CSA) network which can reconstruct the LF with high quality at a low cost of computation.

2 Proposed Method

The proposed SAI-wise and channel-wise attention modules will be elaborated in Section 2.1 and Section 2.2 correspondingly. We adopt the skeleton in [17] as our baseline network in Fig. 1 where the reconstruction network is firstly employed to reconstruct the coarse novel SAIs, and then these intermediate SAIs will be refined in the following refinement network. The stacked layers in the reconstruction network are replaced with U-Net to extract the features from input SAIs. U-Net has been proved to be capable to extract multi-level features carrying hierarchical information [6,7,13]. The proposed modules will be applied to the baseline network to demonstrate their impact.

2.1 SAI-wise Attention

In a LF image, even though all SAIs share a major proportion of information in common, details still vary in different SAIs, especially when occlusion happens. Occlusion often plays a challenging role to cause artifacts. Therefore, paying more attention to the SAIs with more details can be beneficial. To this end, we

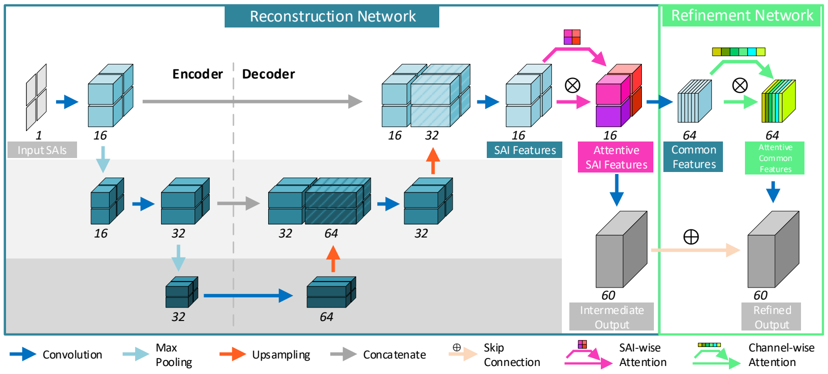

Fig. 1. The illustration of the proposed network architecture. Layers with SAI-wise and channel-wise attention are colored in purple and green respectively.

propose SAI-wise attention as shown in Fig. 2(a). For simplicity, some layers are omitted.

Let x be the 4D input tensor of size (U, V, W, H, C) where U and V are the angular dimensions, W and H are the spatial dimensions and C is the size of the channel space. At first, the dimensions (W, H, C) are shrunk into one dimension by global average pooling to get SAI-wise statistic z^{SAI}, in which s-th element is calculated as

$$z_s^{SAI} = f_{GAP}(x_s) = \frac{1}{W \times H \times C} \sum_i^W \sum_j^H \sum_k^C x_s(i, j, k) \tag{1}$$

where x_s denotes the corresponding SAI. Afterwards, $z^{SAI} \in \mathbb{R}^{(U \times V)}$ is vectorized and fed to three fully connected (FC) layers with n_1, n_2 and $U \times V$ neuron units subsequently, the first two FC layers are followed by ReLU [11] functions and the third one is followed by Sigmoid gating function. Formally, SAI-wise attention is obtained as

$$s^{SAI} = Sigmoid(\mathbb{W}_3 \cdot ReLU(\mathbb{W}_2 \cdot ReLU(\mathbb{W}_1 \cdot z^{SAI}))) \tag{2}$$

where $\mathbb{W}_1$, $\mathbb{W}_2$ and $\mathbb{W}_3$ are trainable parameters of the three FC layers. Finally, the attentive features of $\hat{x}_s$ are obtained as

$$\hat{x}_s = s_s^{SAI} \cdot x_s \tag{3}$$

where $\hat{x}_s$ and s_s^{SAI} indicate the s-th elements of $\hat{x}$ and s^{SAI}.

2.2 Channel-wise Attention

Most of the previous deep learning-based LF methods treat the channels equally, which hinders the feature representation. In the refinement network, the features

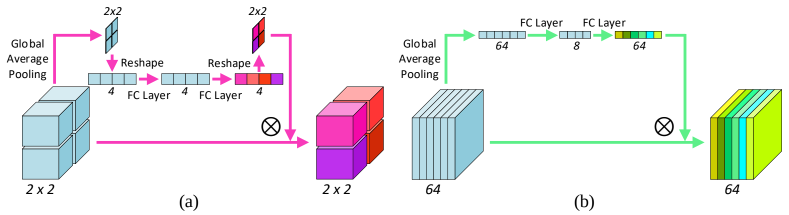

Fig. 2. The illustration of (a) SAI-wise attention and (b) Channel-wise attention.

are convoluted from 4D to 3D which can help to further extract common features and refine the reconstructed SAIs. Inspired by [8], we presume that the channels of the learned common features are not always informative for refinement in some cases and introduce the squeeze-and-excitation channel-wise attention module for our LF reconstruction which is shown in Fig. 2(b).

Given x indicating the 3D input tensor, each channel is denoted as $x_c \in (W, H, F)$ where W and H are spatial width and height and F is the vectorized SAI dimension. A global average pooling layer f_{GAP} is operated on x to get channel-wise statistic $z^{channel} \in \mathbb{R}^C$. Formally, c-th element of $z^{channel}$ is calculated as

$$z_c^{channel} = f_{GAP}(x_c) = \frac{1}{W \times H \times F} \sum_i^W \sum_j^H \sum_k^F x_c(i,j,k). \tag{4}$$

Then, the shrunk features are fed into two FC layers in succession to calculate the attention weights as

$$s^{channel} = Sigmoid(\mathbb{W}_E \cdot ReLU(\mathbb{W}_S \cdot z^{channel})) \tag{5}$$

where the first FC layer $\mathbb{W}_E$ squeezes the features by ratio r and the second FC layer $\mathbb{W}_S$ serves as excitation from the C/r neuron units back to the C neuron units. ReLU and Sigmoid gating functions are applied to the two FC layers respectively. $s^{channel} \in \mathbb{R}^C$ is the set of weights which will be multiplied with x to obtain the attentive features

$$\hat{x_c} = s_c^{channel} \cdot x_c \tag{6}$$

where $\hat{x_c}$ and $s_c^{channel}$ are the c-th channel of the attentive features $\hat{X}$ and channel-wise attention weights $s^{channel}$ respectively.

3 Experiments

3.1 Implementation and Evaluation Details

We implement our proposed CSA method using the deep learning library Keras [3] with Tensorflow [2] backend. To evaluate the proposed method, the experiments have been carried out on extensive LF datasets *30 Scenes* [10] and *Occlusions* [1] which are captured with the Lytro Illum camera. The *30 Scenes* and

Occlusions dataset have 30 and 43 LF images which have no sample overlapped with the training set. The comparison is conducted over (2×2) to (8×8) task and the metrics of Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). With regard to the attention modules, a SAI-wise attention module is applied to the last layer of U-Net before intermediate output with $n_1 = 4$ and $n_2 = 4$. Moreover, a Channel-wise attention module is applied to the second intermediate layer of the refinement network with a ratio of $r = 8$.

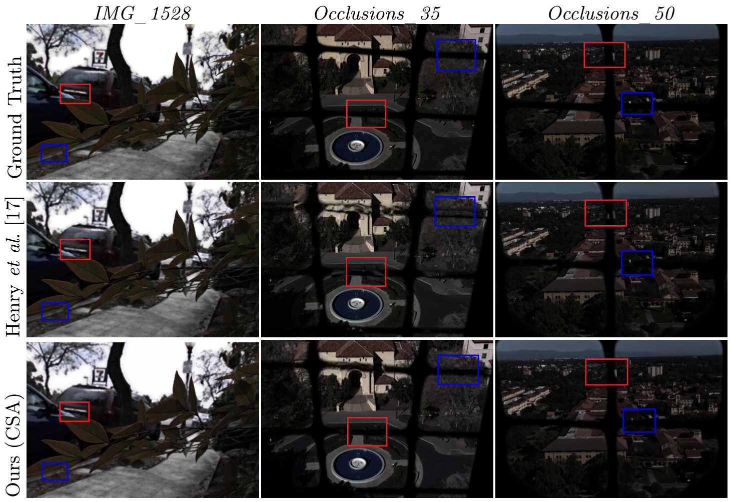

Fig. 3. Visualization of reconstruction.

Table 1. Comparison of overall performance. PSNR and SSIM scores are demonstrated as the metrics of reconstruction quality. Bold scores indicate the best results.

Method	*30 Scenes*	*Occlusions*
Kalantari et al.	37.97/0.9725	31.70/0.8915
Yeung et al.	38.85/0.9759	32.52/0.9029
Ours (CSA)	**39.03/0.9762**	**33.15/0.9067**

3.2 Comparison with State-of-the-art

We compare our method with 2 state-of-the-art methods, Kalantari *et al.* [10] and Yeung *et al.* [17], and the results are shown in Table 1. In the *30 Scenes* dataset, our method achieves the best result outperforming Yeung *et al.* by 0.18 db PSNR. The edge extends to more than 0.50 db in *Occlusions* dataset which

Table 2. Comparison of number of trainable parameters and running speed.

Method	# Trainable Parameters	Speed(Seconds per sample)
Kalantari et al.	1644204	593.47
Yeung et al.	1498752	30.23
Ours (CSA)	**719198**	**13.16**

features the challenging scenarios of occlusion. To demonstrate the algorithm efficiency, we compare the number of trainable parameters and the running speed in Table 2. The running speed is measured by executing the methods in CPU-only mode. It is observed that our method has achieved better performance with a substantially smaller model and higher speed. The method of Yeung *et al.* suffers from the giant network with the 16 stacked alternating filters while our method comes with less than half of the size and runs at 2 times faster speed.

Visualization of some examples is demonstrated in Fig. 3. In *IMG_ 1528*, it is observed that the improvement comes from the area with reflective surfaces, *e.g.* the surface of the vehicle's back in the red box, where the result of Henry *et al.* is light leaking while CSA is reconstructed perfectly. More significant differences are observed in the occluded areas, *e.g.* in *Occlusion_ 35* and *Occlusion_ 50*, the window frames in the red and blue boxes are blurred while CSA reconstructs these parts more completely. In *IMG_ 1528*, the edge occluded by the leaf is also more clear in the result of CSA than the other method.

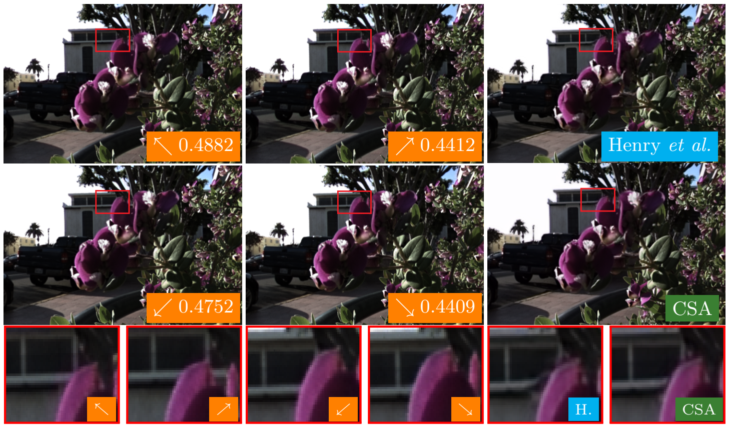

Fig. 4. Visualization of SAI-wise attention on *IMG_ 1555*. The input SAIs placed by (2×2) in the first two columns annotated with attentive weights and arrows indicating their locations. Selected reconstructed SAIs of Henry *et al.* [17] and ours (CSA) are shown in the third column. Selected regions are annotated with the red boxes in the examples and zoomed in the last row.

Table 3. Ablation study of the attention modules.

SAI-wise Attention	Channel-wise Attention	*30 Scenes*	*Occlusions*
✗	✗	38.94/0.9749	32.95/0.9054
✓	✗	38.99/0.9750	33.09/0.9060
✗	✓	39.02/0.9754	33.13/0.9064
✓	✓	**39.03/0.9762**	**33.15/0.9067**

3.3 Study of Attention Modules

In order to investigate the contribution of our proposed attention modules, in this section, an ablation study is conducted by evaluating the model without the attention modules. As shown in Tab. 3, the baseline model without SAI-wise and channel-wise attention modules produces just slightly better performance than Yeung *et al.* [17]. With SAI-wise attention only, 0.05 db and 0.14 db improvement is obtained in *30 Scenes* and *Occlusions* correspondingly. A similar improvement is observed with channel-wise attention solely as 0.08 db and 0.18 db improvement is obtained. Such results demonstrate that the two proposed attention modules are beneficial to the feature representation when they are working separately. If combining these two modules, the performance gains a little bit better by 0.09 db and 0.20 db, meaning these two attention modules have a similar effect on feature representation, hence the improvement has saturated.

3.4 Study of SAI-wise Attention

To further understand how SAI-wise attention module contributes to the performance, we visualize the input SAIs and the corresponding attention weights of selected samples in Fig. 4. It is observed that some SAIs are weighted higher than others. In *IMG_1555*, some details are not occluded in the top left SAI such as the window frame of the house behind the flower annotated by the red box. In the bottom right SAI which is with the lowest weight, the junction of the window is fully occluded. It is possible that treating SAIs equally may cause artifacts because of mixing visible and occluded details. With the feature reinforcement by SAI-wise attention, the window frame is reconstructed by CSA compared to Henry *et al.* which gets distorted content. In terms of the results, SAI-wise attention has successfully learned to weight the SAIs and reinforced the ones with informative details.

4 Conclusion

In this paper, we have presented two attention modules for LF reconstruction which enhance the channel-wise and SAI-wise feature representation respectively. Our experimental results have demonstrated that these two attention modules have succeeded to guides the baseline network to focus on these informative channels and SAIs, and the proposed CSA network can achieve the state-of-the-art performance at a low cost.

References

1. Stanford Lytro Light Field Archive, http://lightfields.stanford.edu/LF2016.html
2. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: 12th $USENIX$ Symposium on Operating Systems Design and Implementation $OSDI$ 16). pp. 265–283 (2016)
3. Chollet, F., et al.: Keras. https://keras.io (2015)
4. Fiss, J., Curless, B., Szeliski, R.: Refocusing plenoptic images using depth-adaptive splatting. In: 2014 IEEE international conference on computational photography (ICCP). pp. 1–9. IEEE (2014)
5. Heber, S., Pock, T.: Convolutional Networks for Shape from Light Field. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3746–3754 (2016)
6. Heber, S., Yu, W., Pock, T.: U-shaped Networks for Shape from Light Field. In: Procedings of the British Machine Vision Conference 2016. vol. 1, pp. 37.1–37.12 (2016)
7. Heber, S., Yu, W., Pock, T.: Neural EPI-Volume Networks for Shape from Light Field. In: Proceedings of the IEEE International Conference on Computer Vision. vol. 2017-Octob, pp. 2271–2279 (Oct 2017)
8. Hu, J., Shen, L., Sun, G.: Squeeze-and-Excitation Networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7132–7141. IEEE (Jun 2018)
9. Jeon, H.G., Park, J., Choe, G., Park, J., Bok, Y., Tai, Y.W., Kweon, I.S.: Depth from a Light Field Image with Learning-Based Matching Costs. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 297–310 (Feb 2019)
10. Kalantari, N.K., Wang, T.C., Ramamoorthi, R.: Learning-based view synthesis for light field cameras. ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2016) **35**(6), 193 (2016)
11. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10). pp. 807–814 (2010)
12. Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., Hanrahan, P.: Light field photography with a hand-held plenoptic camera. Computer Science Technical Report CSTR **2**(11), 1–11 (2005)
13. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
14. Shin, C., Jeon, H.G., Yoon, Y., So Kweon, I., Joo Kim, S.: EPINET: A Fully-Convolutional Neural Network Using Epipolar Geometry for Depth From Light Field Images. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2018)
15. Wang, T.C., Efros, A.A., Ramamoorthi, R.: Occlusion-aware depth estimation using light-field cameras. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3487–3495 (2015)
16. Wang, T.C., Zhu, J.Y., Hiroaki, E., Chandraker, M., Efros, A.A., Ramamoorthi, R.: A 4D light-field dataset and CNN architectures for material recognition. In: European Conference on Computer Vision. pp. 121–138. Springer (2016)
17. Wing Fung Yeung, H., Hou, J., Chen, J., Ying Chung, Y., Chen, X.: Fast Light Field Reconstruction With Deep Coarse-To-Fine Modeling of Spatial-Angular Clues. In: The European Conference on Computer Vision (ECCV) (Sep 2018)